

PHYLOGENOMICS AND SPECIES DELIMITATION IN THE BRUSH-TAILED MOUSE  
GENUS *CALOMYSCUS* USING DDRADSEQ DATA

An Undergraduate Research Thesis

Presented to  
The Department of Evolution, Ecology, and Organismal Biology  
In Partial Fulfillment of the Requirements for Graduation with Research Distinction in Zoology  
The Ohio State University

By  
Brooke Rawson  
May 2019

Project Advisors: Dr. Andreas S. Chavez, Department of Evolution, Ecology and Organismal Biology; Dr. Ryan W. Norris, Department of Evolution, Ecology and Organismal Biology

## Abstract

Brush-tailed mice of the genus *Calomyscus* have been found throughout the varying landscapes across Southwest Asia. Morphological similarity amongst species in conjunction with a lack of sufficient molecular data has led to much uncertainty about the taxonomic relationships among the individual species within the genus. The number of species as well as their evolutionary relationships are not well defined. The objective of this study was to use genome-wide molecular data to discern the historical evolutionary associations of several semi-isolated *Calomyscus* populations located in Pakistan and Southern Iran. We found that there is statistical support for the presence of four distinct species within this region, though the cutoffs for each species remain somewhat ambiguous. Analyses from several bioinformatic programs supported the existence of two genetically distinguishable groups within the presently recognized *C. baluchi* species. The current division between *C. hotsoni*, *C. cf. bailwardi*, and *C. baluchi* was also well supported. The varying elevation and drastic fluctuations in geographic landscapes spanning across the region provide a possible driving force for the divergence and diversification of these populations.

## Introduction

The recent development of inexpensive genome-wide data collection techniques for establishing evolutionary relationships has allowed for a large-scale, cost-effective approach to population-level evolutionary biology and systematics studies. The ability to obtain and compare single nucleotide polymorphism (SNP) data across entire genomes —rather than sequence data from a few select regions— opens up boundless opportunities for extensive phylogenetic comparisons between and within species. Previous studies have demonstrated the increased accuracy and more comprehensive understanding obtained by utilizing genome-wide approaches (Spinks et al. 2014; Yang et al. 2017). Genome-wide SNP data is more cost effective and easier to obtain and store than whole-genome data, without sacrificing the quality of information acquired from it, making it the more practical choice for many taxonomic studies (Dupuis et al. 2012).

Despite being relatively common and widespread throughout southwestern Asia, the taxonomic relationships among members of the rodent genus *Calomyscus* are tenuous and have received limited attention with regards to molecular data (Norris et al 2008; Akbarirad et al. 2016). Morphologically, there is nominal variation between populations, however genetic data suggests that allopatry has led to the formation of multiple distinctive species (Norris et al. 2008). To date, eight geographically isolated species are recognized within *Calomyscus*, which collectively have been found to inhabit a wide variety of landscapes across Syria, Iran, Azerbaijan, Turkmenistan, Afghanistan, and Pakistan (Norris et al. 2008; Kilpatrick 2017). Primarily due to partially conflicting data, the species boundaries of these groups are not yet clearly defined (Esmaeili-Rineh et al. 2007; Norris et al. 2008). Since its initial description by Thomas in 1905, the status of the genus has been constantly debated and revised. Several earlier

descriptions listed only one species, *C. bailwardi*, though many revisions have since been made (Norris et al. 2008). In 2005, Musser and Carleton proposed the presence of the eight distinct species that are currently recognized (Figure 1; Kilpatrick 2017).

Two of the eight species, *C. baluchi* and *C. hotsoni*, are distributed throughout Pakistan. The range of *C. baluchi* includes the mountainous regions of North-central Pakistan and extends into southern Afghanistan, whereas *C. hotsoni* primarily inhabits Western Pakistan and Southeastern Iran. There is some conflicting data on the taxonomic status and species boundaries of these two groups, as well as some speculation that *C. baluchi* may actually be comprised of two distinct species (Norris 2009). Members of *C. bailwardi* have been found across Iran from the northwestern region down to the beginnings of the *C. hotsoni* range in the southeast (Shahabi et al. 2010). Akbaribad et al. (2016) argued that *C. bailwardi* as currently recognized comprises two species, which they identify as *C. bailwardi* and *C. cf. bailwardi* “species Group B.”

The objective of this study was to utilize a genome-wide approach to explore the evolutionary history of these groups. Next-generation restriction site associated DNA sequencing (RADseq) is one such approach that has proven useful in previous systematics and phylogenetic studies (Ree and Hipp 2015; Bateman et al. 2018). Presumed members of the suspected northern and southern populations of *C. baluchi*, as well as *C. hotsoni* and *C. cf. bailwardi* (from “species Group B” of Akbaribad et al. 2016) were represented in the sampling. Using SNP data generated with a RADseq approach, the evolutionary relationships among these species were further disentangled.

## Methods

### *Sample Collection*

Thirty-four tissue samples were used in total (Figure 2). Four were fresh tissue samples collected by J. Darvish, a collaborator in Ferdowsi University of Mashhad in Mashhad, Iran. One sample was collected near Swat, Pakistan and obtained from the Field Museum in Chicago. The remainder of the samples were collected in a series of expeditions between 1991 and 1999 conducted by Charles A. Woods and C. William Kilpatrick, and stored at the University of Vermont in 95% ethanol.

### *DNA Extraction, Size Selection, and Sequencing*

DNA was extracted from the tissues of 34 individuals using a Qiagen DNeasy kit and quantified using a High Sensitivity DNA assay on a Qubit Fluorometer. DNA was digested and size selected according the RADseq protocol described by Peterson et al. (2012), with some modifications. Restriction enzymes *MspI* and *SbfI* were used for the double digest, with *MspI* being the common cutter and *SbfI* the rare cutter. DNA fragments were ligated with barcoded Illumina adapters, the samples were then pooled by eight samples, size-selected for fragments in the range of 300–400 base pairs (bp) using the BluePippin size selector, ligated with Illumina multiplexing indices, and final pools were assessed for quality using Agilent TapeStation electrophoresis. All pools were combined and sequenced for 50bp single-end reads on a single lane on an Illumina HiSeq4000 at the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 OD018174 Instrumentation Grant.

### *Bioinformatics and Data Analysis*

Raw Illumina data were first demultiplexed using the unique combination of barcodes and indices, and filtered using iPyrad 0.7.8 (Eaton 2014), according to the default filters with several modifications. Alterations were made in the params file at lines 9, 10, 14, 21, 22, and 23. At line 9, a maximum of 4 low-quality base pairs was allotted per read to minimize sequence uncertainty. Lines 11 and 12 were set to a minimum depth of 10, to set the lower bound of the read depth at which statistical and majority base calls were made, respectively. Line 14 was increased to a value of 0.95 from the default of 0.90 in order to increase the percent sequence similarity required for sequences to be collapsed into a single cluster; this decreases the chances of falsely calling similar sequences the same sequence. The minimum number of samples that must have data at a given locus for it to be retained in the final data set was set at 97% (33 out of 34 samples: line 21 in params file). Lines 22 and 23 were both assigned values 4 to control the maximum number of SNP's and indels allowed in a locus, respectively.

The number of genetic population groupings was estimated with a 668 SNP dataset from iPyrad using a Bayesian clustering method in STRUCTURE 2.3.4 (Pritchard et al. 2000). The number of genetic clusters was analyzed with a 10,000 burnin period followed by 100,000 reps for each K level (1-7). We processed runs in the CLUMPAK online web server (Kopelman et al. 2015) and evaluated the optimal K following Evanno et al. (2005).

Species delimitation was performed with an 839 loci dataset from iPyrad using BPP 3.4.2 (Flouri et al. 2019) for joint species-delimitation and species-tree analysis following the default filters described by Yang (2015), with some modifications. The BPP program uses a rjMCMC algorithm (Rannala and Yang 2013) to estimate a posterior distribution of species delimitations and species trees, while integrating over uncertainty in gene trees. The assignment of individuals to populations or putative species is required and we assigned individuals to putative species

based on our population-assignment results from the STRUCTURE analysis. This led to grouping individuals into four predefined groups: ((A) *C. baluchi* group; (B) *C. cf. baluchi* group; (C) *C. cf. balwardi*; and (D) *C. hotsoni* group. The  $\theta$  (population-size) parameter was assigned an inverse-gamma prior (3 0.004) and  $\tau$  (divergence time for the root in the species tree) parameter was assigned an inverse-gamma priors (3 0.002). We performed multiple runs with different starting-tree topologies to ensure consistency of the species delimitation analyses. All analyses were run for a total of 20,000 iterations (sampling interval of 2) with a burn-in of 2000.

Phylogenomic relationships among individuals was inferred with a nexus file containing 34 taxa and 35,342 characters using SVDQUARTETS 1.0 (Chifman and Kubatko 2014) implemented in PAUP\* 4.0b10 (Phylogenetic Analysis Using Parsimony; Swofford 2003). We evaluated all possible quartets without prior assignment to populations and used non-parametric bootstrapping with 1,000 replicates for estimating statistical support.

## Results

### *iPyrad*

We obtained a total of 66.52 million sequence reads and an average of 1.96 ( $\pm 1.24$  SD) million reads per individual. All individuals were retained in the final analyses with a range of 1.57–3.13 million reads. On average, the total number of retained loci per individual was 823 ( $\pm 61$  SD).

### *Genetic Clusters*

The STRUCTURE analysis yielded three distinct population clusters (Figure 3). These clusters correspond to populations of northern *C. baluchi*, southern *C. baluchi*, and a third population comprised of both *C. hotsoni* and *C. cf. balwardi* individuals.

### *Species Tree and Species Delimitation*

The results of the BPP species tree and species delimitation analysis found strong support for the presence of four genetically distinguishable groups. These groups were clusters of individuals from populations of northern *C. baluchi*, southern *C. baluchi*, *C. cf. bailwardi*, and *C. hotsoni*. In total, three potential species trees were constructed based on 20,000 runs, all of which followed a four-population scenario (Table 1).

The results of the SVDQUARTETS and BPP species-tree analysis found strong statistical support for *C. hotsoni* and *C. cf. bailwardi* as sister taxa, with all three possible species tree scenarios generated by BPP demonstrating this relationship (Table 2).

### **Discussion**

The plummeting costs and increasing speeds at which genome-wide data can be obtained has made systematics easier, more efficient, and more accurate (Spinks et al. 2014). Multi-locus comparisons across species have proven effective in uncovering evolutionary relationships and constructing phylogenies that were previously considered ambiguous (Chifman and Kubatko 2015). The intention of this study was to use this type of data to progress the overall understanding of a widely understudied group of organisms and contribute to our knowledge of the evolutionary relationships among these populations.

Speculation about the species boundaries within the genus *Calomyscus*, particularly those groups inhabiting Pakistan and Iran, was addressed using multi-locus data analyzed using a variety of bioinformatic programs. BPP and STRUCTURE analyses found different numbers of genetic groups, with BPP finding four species consisting of *C. hotsoni*, *C. cf. bailwardi*, a



northern *C. baluchi* group, and a southern *C. baluchi* group, and STRUCTURE finding three clusters consisting of a northern *C. baluchi* group, a southern *C. baluchi* group, and a third cluster consisting of individuals from both *C. cf. bailwardi* and *C. hotsoni*. The phylogeny constructed via SVD quartets shows strong nodal support (94%) for the divergence between *C. hotsoni* and *C. cf. bailwardi*, which is further supported by the BPP species-tree results (Fig 3 and Table 2). There is remaining ambiguity regarding the precise cutoff between *C. hotsoni* and *C. cf. bailwardi*, however, it is worth noting that the *C. hotsoni* individual that grouped closely with the *C. cf. bailwardi* clade in Figure 3 (NOR22B) was taken from a location in Iran near the collection point of the *C. cf. bailwardi* samples, providing a possible explanation for the genetic similarities between this individual and the members of a separate species.

Both BPP and STRUCTURE supported the existence of two genetically distinct *baluchi* groups. Previous molecular and morphological research has identified a suspected *C. baluchi* subspecies, called *C. b. mustersi*, which primarily inhabits Afghanistan and parts of Northwest Pakistan and may have genetic overlap with the putative northern *C. baluchi* clade identified in this study (Akbarirad et al. 2018; Ellerman 1948). *Calomyscus b. mustersi* may be a subset of the northern *C. baluchi* group, or it could potentially be a *C. baluchi* subspecies distinct from both the northern and southern *C. baluchi* groups. The Kurram River in Northwest Pakistan may serve as a barrier between these groups and thus a driver for their divergence (Norris 2009).

The geographic complexities of Pakistan and the surrounding regions in Southwest Asia provide a unique opportunity to study the effects of varying topography on geographic isolation and speciation. The southern portion of the Hindu Kush mountain range extends vertically through Central Afghanistan and Northern Pakistan and is flanked on either side by relatively flat areas consisting of both deserts and plains. The distribution of *Calomyscus* throughout such a

region has undoubtedly shaped the evolutionary relationships among species. Partial geographic isolation provides an avenue for members of the genus to separate and accumulate sufficient genetic differences to be declared independent species, yet it may also allow for admixture between these species (Payseur and Rieseberg 2016). Any resulting hybrids have the potential to complicate attempts to classify the taxonomy of these groups (McDade 1992).

This study, along with many others that have focused on *Calomyscus*, was primarily limited by restricted access to specimens. The uncertainty surrounding the evolutionary relationships in *Calomyscus* will likely persist until more molecular data are accumulated, though studies of this nature continue to expose progressively more detail about the evolutionary history of the genus. Classifying and determining phylogenetic relationships between species is fundamental to the field of evolutionary biology and is a crucial first step in being aware of and preserving biodiversity (Wheeler 1995). The findings of this research will ideally provide valuable information that can later be used for further study and analysis.

## Acknowledgements

I would like to thank both of my project advisors, Dr. Andreas Chavez and Dr. Ryan Norris, for their valuable support and encouragement throughout this process. I would also like to thank Ola Balkowiec for assistance with DNA extractions.

## Literature Cited

- Akbarirad S., Darvish, J., Aliabadian, M. (2016). Increased species diversity of brush-tailed mice, genus *Calomyscus* (Calomyscidae, Rodentia), in the Zagros Mountains, western Iran. *Mammalia* 80:549-562.
- Akbarirad, S., Darvish, J., Aliabadian, M. (2018). Taxonomic research on *Calomyscus baluchi* from Bamyian in Afghanistan and molecular comparison with *C. baluchi* from Pakistan. *Journal of Research in Biology* 1(1), e1518.
- Bateman, R. M., Sramkó, G., Paun, O. (2018). Integrating restriction site-associated DNA sequencing (RAD-seq) with morphological cladistic analysis clarifies evolutionary relationships among major species groups of bee orchids. *Annals of Botany* 121:85–105.
- Chifman J, and Kubatko L.S. (2014). Quartet inference from SNP data under the coalescent model. *Bioinformatics* 30(23):3317–3324.
- Chifman, J. and Kubatko, L.S. (2015). Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites. *Journal of Theoretical Biology* 374:35-47.
- Dupuis, J. R., Roe A. D., Sperling F. A. (2012). Multi-locus species delimitation in closely related animals and fungi: one marker is not enough. *Molecular Ecology* 21:4422–4436
- Eaton, D.A.R. (2014). PyRAD: assembly of de novo RADseq loci for phylogenetic analyses. *Bioinformatics* 30:844– 1849.
- Ellerman, J.R. (1948). Key to the rodents of southwest Asia in the British Museum collection. *Proc. Zool. Soc. Lond.* 118:765– 816.
- Esmaeili-Rineh S., Darvish, J., Hadad F., Ghasemzadeh, F. (2008). A new karyotype of *Calomyscus* from the Khorasan province, Iran. *Hystrix* 19: 67–71.
- Evanno, G, Regnaut, S, Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Molecular Ecology* 14:2611– 2620.

- Flouri T., Jiao X., Rannala B., Yang Z. (2018). Species Tree Inference with BPP using Genomic Sequences and the Multispecies Coalescent. *Molecular Biology and Evolution* (accepted manuscript).
- Kilpatrick, C.W. (2017). Family Calomyscidae (Brush-tailed Mice) Pp. 144-155 in D.E. Wilson, T.E Lacher, and R.A. Mittermeier. *Handbook of Mammals of the World* Vol. 7. Lynx Edicions.
- Kopelman, N.M., Mayzel, J., Jakobsson, M., Rosenberg, N.A., Mayrose, I. (2015). Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. *Molecular Ecology Resources*, 15:1179– 1191.
- McDade, L.A. (1992). Hybrids and phylogenetic systematics II. The impact of hybrids on cladistic analysis. *Evolution* 46:1329–1346.
- Musser, G. G. and Carleton, M. D. (2005). Subfamily Murinae. In: Wilson, D.E., Reeder, D.M. (Eds.), *Mammal species of the world, A taxonomic and geographic reference*, The Johns Hopkins University Press Baltimore.
- Norris, R.W. 2009. Phylogenetic relationships and divergence times in rodents based on both genes and fossils. Unpublished Ph. D. dissertation, University of Vermont, Burlington.
- Norris, R.W., Woods, C.A., Kilpatrick, C.W. (2008). Morphological and molecular definition of *Calomyscus hotsoni* (Rodentia: Muroidea: Calomyscidae). *Journal of Mammalogy* 89:306-315.
- Payseur, B. A., and Rieseberg, L. H. (2016). A genomic perspective on hybridization and speciation. *Molecular Ecology* 25: 2337– 2360.
- Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., Hoekstra, H. E. (2012). Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species. *PLOS One* 7(5):1-11.
- Pritchard, J. K., Stephens M., Donnelly, P. J. (2000). Inference of population structure using multilocus genotype data. *Genetics* 155:945-959.
- Rannala, R. and Yang, Z. (2013). Improved Reversible Jump Algorithms for Bayesian Species Delimitation. *Genetics* 194:245-253.
- Ree, R. H., Hipp, A. L. (2015). Inferring phylogenetic history from restriction site associated DNA (RADseq) Pp. 181-204 in HÖrandl, E. and Appelhans, M.S. (eds.), *Next-Generation Sequencing in Plant Systematics*.
- Shahabi, S., B. Zarei and B. Sahebjam. (2010). Karyologic study of three species of *Calomyscus* (Rodentia: Calomyscidae) from Iran. *Iranian Journal of Animal Biosystematics* 6:55–60.

- Spinks, P. Q., Thomson, R. C., & Shaffer, H. B. (2014). The advantages of going large: Genome-wide SNPs clarify the complex population history and systematics of the threatened western pond turtle. *Molecular Ecology* 23:2228– 2241.
- Swofford, D. L. (2003). PAUP\*. Phylogenetic Analysis Using Parsimony (\*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.
- Thomas, O. (1905). On a collection of mammals from Persia and Armenia presented to the British Museum by Col. A. C. Bailward. *Proceeding of Zoological Society of London*, 1905 (2):519-527.
- Wheeler, Q.D. (1995) Systematics, the scientific basis for inventories of biodiversity. *Biodiversity and Conservation* 4:476– 489.
- Yang, Z. (2015). The BPP program for species tree estimation and species delimitation. *Current Zoology* 61(5):854–865.
- Yang, B., Cui, L., Perez-Enciso, M., Traspov, A., Crooijmans, R., Zinovieva, N., Megens, H. J. (2017). Genome-wide SNP data unveils the globalization of domesticated pigs. *Genetics, selection, evolution* 49:71.

## Figures

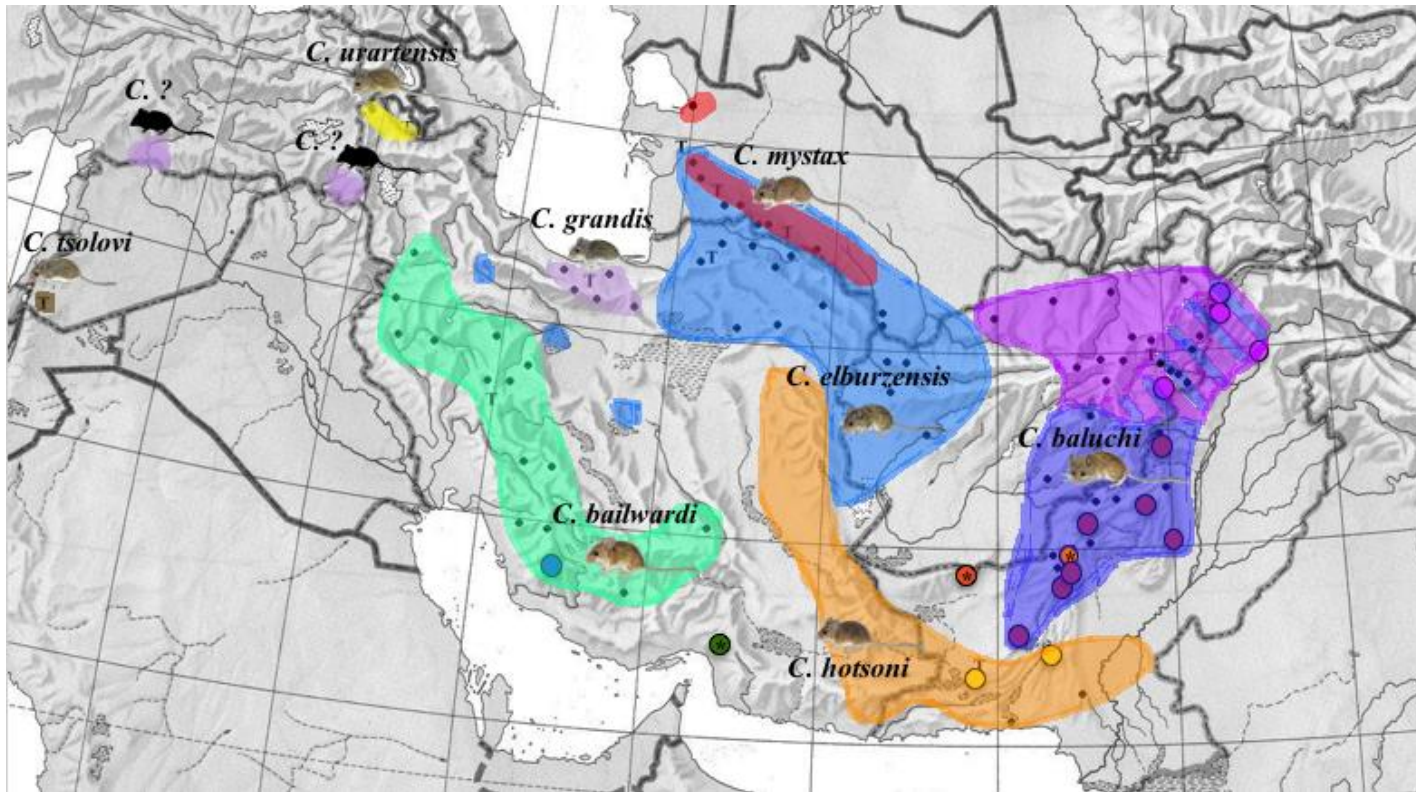


Figure 1: *Calomyscus* species distribution. *C. baluchi* range from Afghanistan to Pakistan may be comprised of both a northern and southern *C. baluchi* species, represented by the division between the pink and purple regions, respectively. Like Kilpatrick (2017), this map does not differentiate the two potential species within *C. bailwardi* suggested by Akbarirad et al. (2016). Due to limited sampling in Southern Iran, *C. bailwardi* and *C. hotsoni* ranges may come closer than depicted.

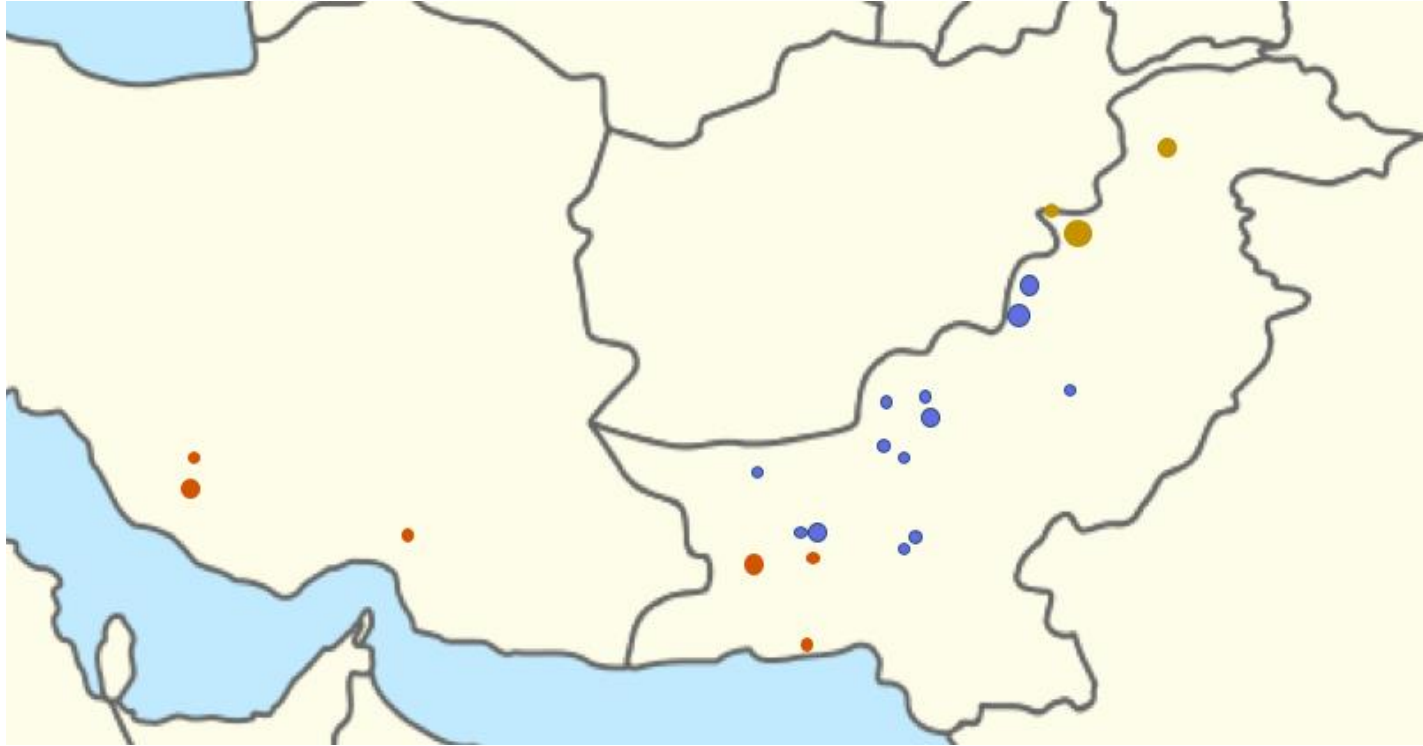


Figure 2: Sample distribution. Colors correspond to STRUCTURE populations: blue represents southern *C. baluchi*, yellow represents northern *C. baluchi*, orange represents *C. hotsoni*/cf. *bailwardi* group.

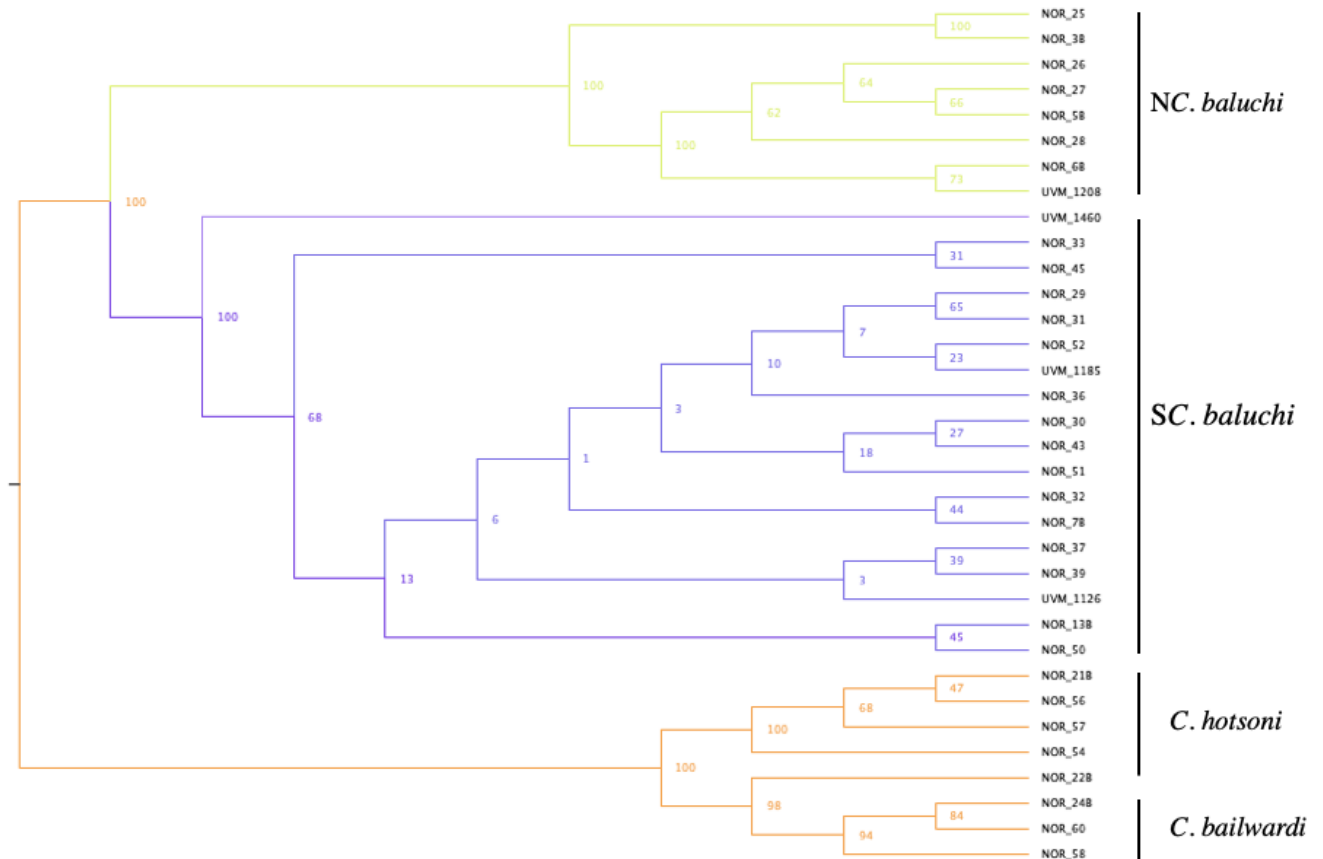


Figure 3: Species tree and species delimitation results. Branches are color coded by STRUCTION results: blue represents southern *C. baluchi*, yellow represents northern *C. baluchi*, orange represents *C. hotsoni*/cf. *bailwardi* group. Vertical bars represent BPP species delimitation results where NC. *baluchi* denotes northern *C. baluchi*, SC. *baluchi* represents southern *C. baluchi*, and *C. hotsoni* and *C. cf. bailwardi* remain distinct groups.



## Tables

Table 1: BPP species delimitation results. N *C. baluchi* represents individuals from the northern *baluchi* population and S *C. baluchi* denotes individuals from the southern *baluchi* group.

Number of Runs	Species Delimitation	Posterior Probability
20000	(N <i>C. baluchi</i> , S <i>C. baluchi</i> , <i>C. cf. bailwardi</i> , <i>C. hotsoni</i> )	1.000000

Table 2: BPP species tree models and their probabilities. N *C. baluchi* represents individuals from the northern *baluchi* population and S *C. baluchi* denotes individuals from the southern *baluchi* group.

Species Tree Model	Posterior Probability
((N <i>C. baluchi</i> , S <i>C. baluchi</i> ), ( <i>C. hotsoni</i> , <i>C. cf. bailwardi</i> ))	.8884
((N <i>C. baluchi</i> , (S <i>C. baluchi</i> , ( <i>C. hotsoni</i> , <i>C. cf. bailwardi</i> ))))	.0843
((S <i>C. baluchi</i> , (N <i>C. baluchi</i> , ( <i>C. hotsoni</i> , <i>C. cf. bailwardi</i> ))))	.0274

Table 3: Prior and posterior probabilities for number of species.

Number of Species	Prior Probability	Posterior Probability
1	.238095	0.000000
2	.238095	0.000000
3	.285714	0.000000
4	.238095	1.000000